An Adaptive Neuro-Fuzzy Inference System Based Approach to Real Estate Property Assessment

Authors

Jian Guan, Jozef Zurada, and Alan S. Levitan

Abstract

This paper describes a first effort to design and implement an adaptive neuro-fuzzy inference system-based approach to estimate prices for residential properties. The data set consists of historic sales of houses in a market in the Midwest region of the United States and it contains parameters describing typical residential property features and the actual sale price. The study explores the use of fuzzy inference systems to assess real estate property values and the use of neural networks in creating and fine-tuning the fuzzy rules used in the fuzzy inference system. The results are compared with those obtained using a traditional multiple regression model. The paper also describes possible future research in this area.

In real estate property value assessment, a multitude of features/attributes are often used to determine a property's fair value. Traditionally such property value assessments have often been conducted with multiple regression-based methods. Recent years have seen an increasing interest in the use of non-conventional methods for property value assessment. The most commonly studied non-conventional methods are neural network-based (Do and Grudnitski, 1992; Worzala, Lenk, and Silva, 1995; Guan and Levitan, 1996; McGreal, Adair, McBurney, and Patterson, 1998; Connellan and James, 1998; Bee-Hua, 2000; and Nguyen and Cripps, 2001). The main motivation in the use of neural networks in property value assessment is the ability of such methods to interpret the numerous and complex interactions of common attributes in a typical property. Though fuzzy logic has also been proposed as an alternative method (Byme, 1995; and Bagnoli, Smith, and Halbert, 1998), there has been no empirical study that uses a fuzzy logic based approach to assess real estate property values.

This paper describes the design and implementation of an adaptive neuro-fuzzy inference system-based approach to estimate prices for residential properties. The data set consists of historic sales of houses in a market in the Midwest region of the United States and it contains parameters describing typical residential property

features and the actual sale price. The study explores the use of fuzzy inference systems to assess real estate property values and the use of neural networks in creating and fine tuning the fuzzy rules used in the fuzzy inference system.

Background

Recent years have seen considerable interest in the use of non-conventional methods in real estate property assessment. The most commonly studied such methods are neural networks-based based (Do and Grudnitski, 1992, 1993; Worzala, Lenk, and Silva, 1995; Guan and Levitan, 1996; McGreal, Adair, McBurney, and Patterson, 1998; Connellan and James, 1998; Bee-Hua, 2000; and Nguyen and Cripps, 2001). The appeal of neural network-based methods lies in the fact that they do not depend on assumptions about the data (e.g., normality, linearity) and may better replicate a home buyer's heuristic thought processes (Guan and Levitan, 1996).

A fuzzy logic framework has also been proposed as an alternative to conventional property assessment approaches (Byme, 1995; and Bagnoli, Smith, and Halbert, 1998). Byme discusses the applicability of fuzzy logic in real estate analysis and contends that fuzzy logic has value as a tool for dealing with the risk and uncertainty in real estate analysis. Bagnoli et al. examine how fuzzy logic may be used in expressing the inherent imprecision in the way that people think and make decisions about the pricing of real estate. Bagnoli et al. believe that the estimated selling price produced by a fuzzy system should be more realistic than that produced by linear regression. Both Byme and Bagnoli et al. have pointed to the potential of fuzzy logic in property assessment. To the best of the authors' knowledge, however, there has been no reported actual study on the use of fuzzy inference systems in real estate assessment.

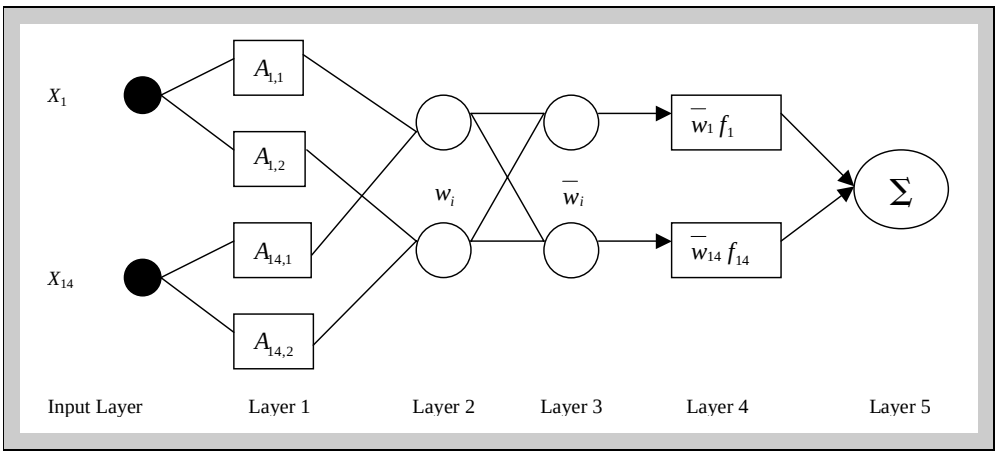
Fuzzy logic has been widely studied in the relatively ''hard sciences'' such as different fields of engineering with varying degrees of success (Ponnambalam, Karray, and Mousavi, 2002; Stepnowski, Mosynski, and Dung, 2003; and Hasiloglu, Yilmaz, Comakli, and Ekmekci, 2004). Fuzzy inference systems (FIS), which are based on fuzzy logic, often consist of IF-THEN rules that fire in parallel when the IF conditions are met. The consequents of rules fire to the degree to which the antecedents of the rules are met. One of the main challenges of creating an FIS is the determination of fuzzy sets and fuzzy rules. Determination of such fuzzy sets and rules requires deep domain knowledge from human experts and the fine tuning of the fuzzy rules and sets can be very time consuming. A solution to this problem is to combine the advantages of a fuzzy system with the learning capability of artificial neural networks (Jang, 1993). The result is an adaptive neuro-fuzzy inference system (ANFIS). ANFIS has the features of the neural networks such as learning abilities, optimization abilities, and the fuzzy inference system such as human-like reasoning using IF-THEN rules and ease of incorporating human expert knowledge. IF-THEN rules can be created (learned) and refined from preexisting data sets.

ANFIS-based systems have been studied in many different fields. Stepnowski, Mosynski, and Dung (2003) apply ANFIS to seabed characterization and find neuro-fuzzy techniques to be a promising tool. Hasiloglu, Yilmaz, Comakli, and Ekmekci (2004) discuss the use of ANFIS in predicting transient heat transfer and find that ANFIS produces better results when compared with methods based on multiple regression analysis. Ponnambalam, Karray, and Mousavi (2002) use ANFIS-based techniques to provide the trajectory of optimal releases and storage reservoir for simulated stochastic inflows and find that ANFIS-based methods perform much better than regression methods. Real estate property assessment seems to offer an ideal setting for applying an ANFIS-based approach where there is a preexisting data set of property features and sale prices. In residential real estate, such historic data are not difficult to obtain. Like neural networks, an ANFIS-based system has an advantage over traditional statistical estimation as it does not require a mathematical modeling of the data set. Given the consensus that fuzzy logic-based systems offer promise to the field of property value assessment (Byme, 1995; and Bagnoli, Smith, and Halbert, 1998) and its success outside of the field of property assessment, the use of ANFIS in property assessment is certainly worthy of further investigation.

Adaptive Neuro-Fuzzy Inference System

The ANFIS study presented in the paper is based on the Sugeno fuzzy model, which was proposed in an effort to develop a systematic approach to generating fuzzy rules and membership function parameters for fuzzy sets from a given input–output data set (Sugeno and Kang, 1988; and Jang, 1993). Exhibit 1 shows the architecture of the ANFIS model. The ANFIS architecture has two sets of trainable parameters: the antecedent membership function parameters, or

Exhibit 1 | The ANFIS Architecture



antecedent parameters, and the polynomial parameters (consequent parameters). The ANFIS architecture uses a gradient descent algorithm to optimize the antecedent parameters and a least squares algorithm to solve for the consequent parameters. Each rule in the ANFIS model is of the form:

$$\begin{aligned} &IF\ x_1\ is\ A_{1,j}\ AND\ x_2\ is\ A_{2,j}\ AND \cdots AND\ x_n\ is\ A_{n,j} \\ &THEN\ y = c_0 + c_1x_1 + c_2x_2 + \cdots + c_nx_n, \end{aligned} \tag{1}$$

where $A_{i,j}$ is the jth linguistic term (such as small, large) of the ith input variable x_i and n is the number of inputs (14 in the current paper). y is the estimated sale price. c_i are consequent parameters to be determined in the training process. Since each rule has a crisp output (by contrast to the fuzzy output), the overall output is obtained via a weighted average.

As shown in Exhibit 1, Layer 0 represents the input layer. The layer has 14 nodes representing the 14 features of a property (see Exhibit 2). In the Sugeno ANFIS architecture, consecutive layers are dedicated to different tasks, creating a process of gradual refining of the model. The learning process consists of a forward pass and a backward pass. During the forward pass, the antecedent parameters are fixed and the consequent parameters are optimized using a least squares algorithm. The backward pass uses gradient decent algorithm to adjust the antecedent parameters of the membership functions for the input variables. The output is calculated as a weighted average of the consequent parameters. Any output error is then used to adjust the antecedent parameters using a backpropagation algorithm. Details of each layer are given as follows.

Layer 1 represents nodes where each node is a function:

$$\mu_{A_i}(x), \tag{2}$$

where x is the input and A_i is the linguistic label for the *ith* node. $\mu_{A_i}(x)$ is also referred to as the membership function for the node. Membership functions commonly used in ANFIS include Gaussian function such as:

$$\mu_{A_i}(x) = \exp\left(\frac{-(x - c_i)^2}{2\sigma_i^2}\right), \tag{3}$$

where c_i and σ_i are centers and widths of the function and are referred to as the antecedent parameters for the membership function. Another commonly used function is the bell-shaped function:

Exhibit 2 | Sale Record Structure, Sample Sale Record, and Sample Preprocessed Input Record

Fields in a Sale Record	Sample Sale Record	Preprocessed Input Record
Street name[a]	Elm Street	Variable rejected
Street address[a]	421 Elm Street	Variable rejected
Sale price[b]	$68,500 (original)	$107,818 (inflation adjusted)
ID[a]	22040700230000	Variable rejected
Sale date[a]	00/00/83	Variable rejected
Neighborhood	537	1
Lot size[a]	1	Variable rejected
Construction type	3	3
Wall type	1	1
Year built[c]	30	30
Square footage of the basement	0	0
Square footage on the first floor	1,373	1,373
Square footage on the second floor	667	667
Square footage in the upper area	0	0
Number of baths	3	3
Presence of central air	0	0
Number of fireplaces	1	1
Basement type	1	1
Garage type	2	2
Garage size (number of cars)	2	2

Notes:
[a]Five input variables are rejected from analysis.
[b]Sale price is an output variable.
[c]The variable Year built represents the last two digits of the actual year that the house was built. Thus, 30 refers to the year 1930.

$$\mu_{A_i}(x) = \frac{1}{1 + [((x - c_i)/a_i)^2]^{b_i}}, \tag{4}$$

where a_i, b_i, and c_i are the antecedent parameters.

These parameters are adaptive and they determine the value of the *ith* membership function for each variable to the fuzzy set A_i. These functions have values ranging from 0 to 1. As the values of the parameters change, the value of the function varies accordingly, therefore indicating the degree to which the input variable x

satisfies the membership function. In this layer there are $n \times k$ nodes where n is the number of the input variables and k is the number of membership functions. For example, if First Floor Area is an input variable and there are two linguistic labels (membership functions) SMALL and LARGE, then there will be two nodes in this layer for First Floor Area denoting the variable's membership values to the linguistic labels SMALL and LARGE.

Layer 2 provides the strength of each rule by means of multiplication as follows:[1]

$$w_i = \prod \mu_{A_i}(x_j),\ i = 1, 2;\ j = 1,..., n. \tag{5}$$

Every node in this layer is the multiplication of the input values from the previous layer. The resulting value represents the firing strength of the *ith* rule where the variable x_j has linguistic value A_i. For example, assume there are only two input variables, First Floor Area and Wall Type, and their linguistic labels are SMALL and LARGE for First Floor Area and STURDY and STURDIER for Wall Type (Exhibits 3 and 4). Then there are two rules whose antecedent parts are as follows: (1) *If the First Floor Area is SMALL AND Wall Type is STURDY*; and (2) *If the First Floor Area is LARGE AND Wall Type is STURDIER.* Thus the number of nodes in this layer is the same as the number of rules. The nodes in this layer are not adaptive.

Layer 3 is the normalization layer where the rule strength is normalized as follows:

$$\overline{w}_i = \frac{w_i}{\sum w_i}, \tag{6}$$

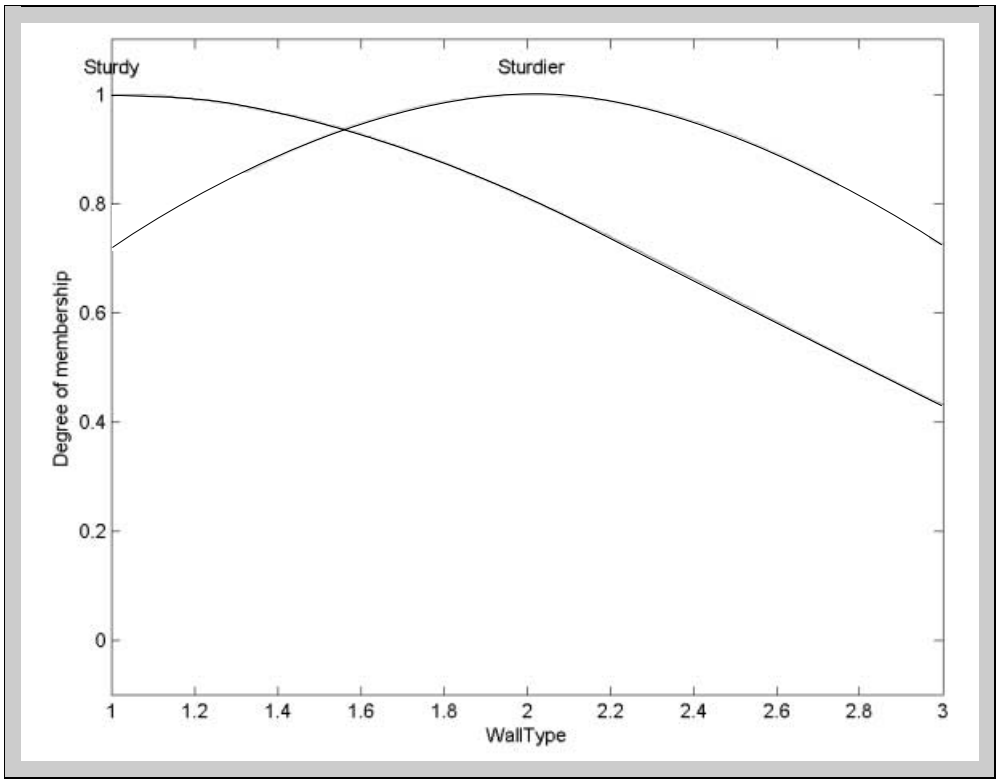
where w_i is the firing strength of the *ith* rule. The number of nodes in this layer is the same as in the last layer and this layer computes each rule's firing strength to the sum of all rules' firing strengths.

Layer 4 is an adaptive layer. Every node in this layer is a linear function and the coefficients of the function are adapted through a combination of least squares approximation and back-propagation.

$$\overline{w}_i f_i = \overline{w}_i(c_0 + c_1x_1 + c_2x_2 + \cdots + c_nx_n). \tag{7}$$

Layer 5 is the output layer. The result of this layer is obtained as a summation of the outputs of the nodes in the previous layer as:

Exhibit 3 | Membership Functions for the WallType Variable
Scenario 1: All Variables, Run 39

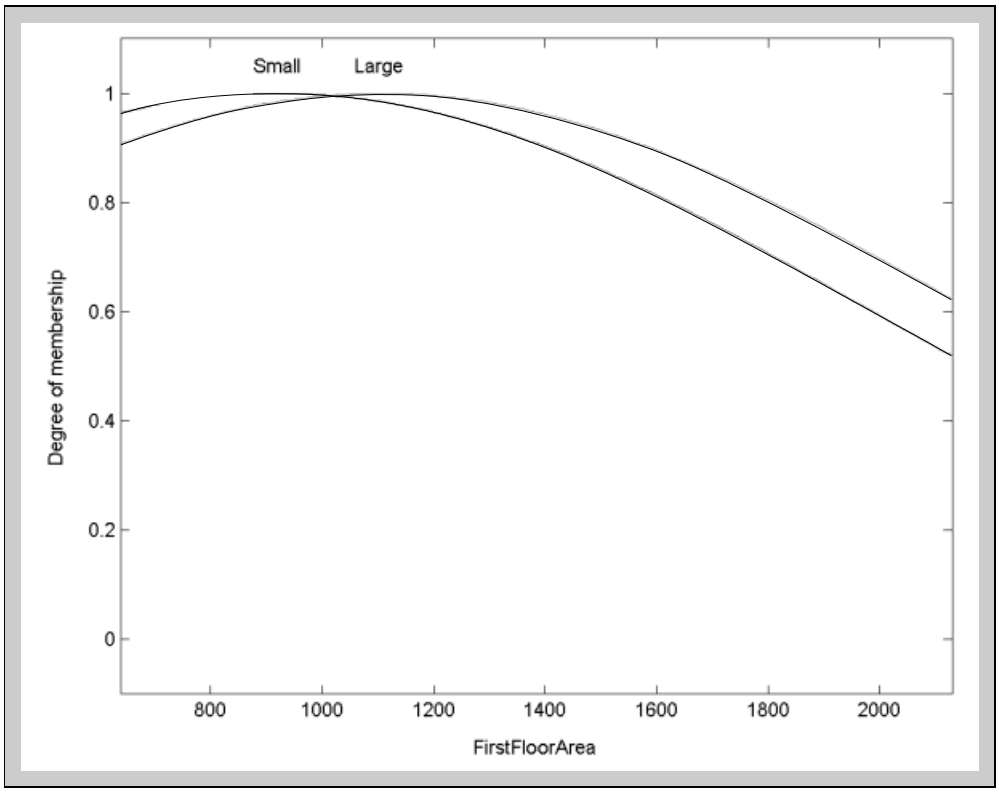


$$\sum_i \overline{w}_i f_i = \frac{\sum w_i f_i}{\sum w_i}, \tag{8}$$

where $\overline{w}_i f_i$ is the output of the node i in the previous layer. The overall output is linear, even though the premise parameters are nonlinear. A numerical example that demonstrates how a predicted sale price is calculated by Layers 1 through 5 is presented in the Results and Discussion Section of the paper.

The ANFIS system thus enables adaptation of the membership functions through fine-tuning of the antecedent parameters and automatically combining different antecedents with different consequents during the evolution of the system. The ANFIS system partitions the multidimensional feature space of properties into several fuzzy spaces and represents the output space with a linear function. The number of partitions in each dimension corresponds to the number of fuzzy

Exhibit 4 | Membership Functions for the FirstFloorArea Variable
Scenario 1: All Variables, Run 39



sets/membership functions in the dimension. The linear combination of the consequents allows approximation of a nonlinear system.

As can be seen from the brief description of the ANFIS architecture above, the main strength of this approach lies in its ability to generate fuzzy rules from a given input-output data set. In simple applications where there are few variables, or a predetermined model structure based on the characteristics of variables is known, or input/output data are not readily available, one has no choice but to build membership functions and fuzzy rules manually using common sense, intuition, and domain knowledge. In more involved cases, such as estimating real estate property values, input/output data are available but the relationships between the variables are complex. In such cases one cannot just look at the data and discern membership functions and fuzzy rules. Given the recognized need for applying fuzzy logic to real estate property valuation (Byme, 1995; and Bagnoli, Smith, and Halbert, 1998), ANFIS naturally offers an attractive opportunity to map the complex and nonlinear relationships between the attributes of a property

and the sale price using automatically-generated fuzzy rules and membership functions.

Data Set Description

The data used in this study consist of 363 single-family house sales in the Midwest region of the U.S. from 1982 to 1992. These sales were from two neighborhoods. The first two columns in Exhibit 2 list the fields in each sale record together with field values of a sample record.

All fields in a property record can potentially serve as input fields except the price. Fourteen of these fields were selected as initial input (see the last column in Exhibit 2). The second column in Exhibit 2 also represents field values of a sale record before it was preprocessed and the corresponding preprocessed values are given in the last column.

The sale prices in the data set are adjusted for inflation before they are used in the study. The following formula was used in adjusting the prices for inflation:

$$\textit{Inflation Adusted}\ \text{Price} = \text{Price} \times \frac{\$88{,}798}{\textit{Average of that Year}},$$

where the average price of each year is as given in Exhibit 5. Exhibit 5 contains the average sales prices for all area houses for each year and was obtained from the county property assessment office. Because any given house will sell for a different (higher) price in a later year than it would in a previous year, even with no change in its attributes, sale prices had to be adjusted for inflation. An alternative would be to limit the sample to sales in a single year. But that itself would limit the validity and generalizability of the results. Another alternative would be to include sale year as an input attribute. But additional dimensionality would reduce the domain coverage and the strength of the model. Thus an adjustment was made for inflation. The numerator \$88,798 is the average price of the last year (1992) in the data set and the denominator is the average price for the year in which the sale occurred. For example, the sample record in Exhibit 2 has an original sale price of \$68,500 and its inflation adjusted price is \$107,818 computed as follows:

$$\$107{,}818 = \$68{,}500 \times \frac{\$88{,}798}{\$56{,}416},$$

where \$56,416 is the average house price for the year 1983 (Exhibit 5) and \$88,798 is the average price of the last year (1992) in the data set.

Exhibit 5 | Inflation Adjustment of Prices

Year	Average Price
1982 and before	$53,092
1983	$56,416
1984	$58,381
1985	$61,376
1986	$62,636
1987	$67,214
1988	$71,699
1989	$76,871
1990	$79,699
1991	$82,892
1992	$88,798

Simulation Architecture

The simulation model is built with the Fuzzy Logic Toolbox of Math Works. An initial input data set contains 363 cases and 19 input variables. In all data analysis projects, errors such as impossible values, impossible combination of values, inconsistent coding, repeated records, outliers, and missing values must be detected, investigated, and corrected (if possible). This is typically done by exploratory graphics and descriptive statistics, such as frequency counts, averages, medians, and minimum and maximum values. The basic descriptive statistics for the original data set, including frequency counts, and the data set used for final analysis are presented in Exhibits 6 and 7. In one instance in the original data set, the size of the basement was coded as -1. The features of each house sold were carefully examined and found that several original sale prices contained inherent errors. For example, in a few instances they were as low as $4,500. Because the goal of the models was to predict the sale prices of typical properties, not outliers, the erroneously coded cases were removed and those cases for which the adjusted sale price was outside $\pm$ 3 standard deviations from the mean adjusted sale price. As a consequence, eight outliers were identified and removed from the data set. As a result, the data set used for analysis contained 355 cases (see Exhibit 7). Filtering extreme values from the data tends to produce better models because the parameter estimates are more stable (McGreal, 1998). Five variables were also removed from the original data: Street Name, Street Address, ID, Year Sold (Sale Date), and Lot size. The Street Name variable is a nominal variable that had 26 different categories and the proper coding for this variable would require 26 additional input variables, which would substantially increase the dimensionality

Exhibit 6 | Values Taken, Frequency Counts, and Percentages for Variables of the Original Data Set Before and After Transformation

	Raw Data Set			Raw Data Set (Status of Variable or Variable Values)		
Variable Name	Values Taken	Frequency	Percent	Values Taken	Frequency	Percent
Sale price	Interval scale			Values not transformed		
Sale price (inflation adjusted)	Same as above			Values not transformed		
Year sold	Same as above			Variable removed (was used in calculation of inflation adjusted price)		
Neighborhood	537	197	54.3	537 coded as 1		
	542	166	45.7	542 coded as 2		
Lot size	1	358	98.62	Variable removed		
	2	4	1.10			
	3	1	0.28			
Construction type	1	269	74.10	Values not transformed		
	2	62	17.08			
	3	32	8.82			
Wall type	1	133	36.64	Values not transformed		
	2	209	57.58			
	3	21	5.79			
Year built	Interval scale			Values not transformed		
Square footage in the basement	Same as above			Values not transformed		
Square footage on the first floor	Same as above			Values not transformed		
Square footage on the second floor	Same as above			Values not transformed		
Square footage in the upper area	Same as above			Values not transformed		
Number of baths	1	227	62.53	1	227	62.53
	2	58	15.98	2	58	15.98
	3	70	19.28	≥3	78	21.49
	4	6	1.65	(Value 4 coded as 3)		
	6	2	0.55	(Value 6 coded as 3)		
Presence of central air	0	103	28.37	Values not transformed		
	1	260	71.63	Values not transformed		
Number of fireplaces	0	62	17.08	0	62	17.08
	1	293	80.72	1	293	80.72
	2	7	1.93	≥2	8	2.20
	3	1	0.28	(Value 3 coded as 2)		

Exhibit 6 | (continued)
Values Taken, Frequency Counts, and Percentages for Variables of the Original Data Set Before and After Transformation

	Raw Data Set			Raw Data Set (Status of Variable or Variable Values)		
Variable Name	Values Taken	Frequency	Percent	Values Taken	Frequency	Percent
Basement type	0	15	4.13	Values not transformed		
	1	21	5.79	Values not transformed		
	2	327	90.08	Values not transformed		
Garage type	0	42	11.57	0	42	11.57
	1	8	2.20	1	8	2.20
	2	297	81.82	2	297	81.82
	3	14	3.86	≥3	16	4.41
	4	1	0.28	(Value 4 coded as 3)		
	5	1	0.28	(Value 5 coded as 3)		
Garage size (number of cars	0	42	11.57	Values not transformed		
	1	251	69.15	Values not transformed		
	2	70	19.28	Values not transformed		

of the data set. In addition, this variable would not likely add to the predictive power of the designed models. Variables Street Address, ID, and Lot Size do not have any predictive ability either. For example, the values stored in the variable Lot Size were identical for all but five of the cases, and variable Year Sold was used to calculate the adjusted sale prices. As a consequence, the data set contained only 14 input variables. To reduce the dimensionality of the data set, less frequently occurring values were also transformed for some variables by assigning them to adjacent categories, as shown in Exhibit 6.

Computer simulations were run for three different scenarios. In the first scenario, all 14 input variables were used. In the second and third scenarios, the PCA and R^2 variable/feature reduction techniques were used. In neural networks one often encounters situations where there are a large number of variables in the data set. In such situations it is very likely that subsets of variables are highly correlated with each other. The accuracy and reliability of a classification or prediction model will suffer if one includes highly correlated variables or variables that are unrelated to the outcome (Mitchell, 1997). One of the key steps in neural networks is finding ways to reduce dimensionality without sacrificing accuracy.

The following two-step process is performed when R^2 variable selection criterion is applied:

Exhibit 7 | Descriptive Statistics for the Variables of the Data Set Used for Analysis

Variable	Average	Std. Dev.	Max	Min	Median
Sale price	$60,613	$12,220	$98,000	$22,500	$59,000
Sale price (inflation adjusted)	$75,459	$14,984	$121,984	$37,494	$74,643
Neighborhood	1.5	0.5	2	1	1
Construction type	1.3	0.6	3	1	1
Wall type	1.7	0.6	3	1	2
Year built	40.7	8.3	90	20	41.0
Square footage in the basement	604.3[a]	266.0[a]	1,394[a]	144[a]	520[a]
	136.2[d]	282.3[d]	1,394[d]	0[d]	0[d]
Square footage on the first floor	1,039.8	179.5	2,130	588	1,027
	794.4[b]	159.3[b]	1,316[b]	528[b]	799[b]
Square footage on the second floor	62.7[d]	218.9[d]	1,316[d]	0[d]	0[d]
	450.3[c]	130.9[c]	913[c]	156[c]	420[c]
Square footage in the upper area	296.8[d]	238.7[d]	913[d]	0[d]	364[d]
Number of baths	1.6	0.8	3	1	1
Presence of central air	0.7	0.5	1	0	1
Number of fireplaces	0.9	0.4	2	0	1
Basement type	1.9	0.4	2	0	2
Garage type	1.8	0.6	3	0	2
Garage size (number of cars)	1.1	0.5	2	0	1

Notes:
[a]Only 80 houses with a basement are included in the calculations.
[b]Only 28 houses with a second floor are included in the calculations.
[c]Only 234 houses with an upper area are included in the calculations.
[d]All 355 houses are included in the calculations. (If a property does not have a basement, second floor or upper area, its feature value is represented by a 0.)

1. Compute the squared correlation for each variable and then reject those variables that have a value less than the cutoff criterion.
2. Evaluate the remaining significant (chosen) variables using a forward stepwise regression. Reject variables that have a stepwise R^2 improvement less than the cutoff criterion (www.sas.com).

Principal component analysis (PCA) is a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components. The objective of principal component analysis is to reduce the dimensionality (number of variables) of the dataset but retain most of the original variability in the data. The first principal component

accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible.

In the R^2 method, the original 14 input variables were reduced to 11 variables (Year Built, Area on the Second floor, and Number of Baths were removed.) The defaults such as squared correlation cutoff = 0.005 and stepwise R^2 improvement cutoff = 0.0005 were used. In the PCA method, the 14 input variables were reduced to only four principal components. The four principal components accounted for 99.98% of the total variation in the data set.

A common approach in neural network applications is to randomly partition the data set into two non-overlapping sets: training set (70% of cases) and validation set (30% of cases) and run computer simulation. Giudici (2003), however, cautions that this approach may produce results that are too optimistic. Therefore, the input data was partitioned into the training, validation, and test data sets. As a result, 40% (142 cases), 30% (107 cases), and 30% (106 cases) of the data set were randomly allocated to the training, validation, and test subsets, respectively. The training set is used for preliminary model fitting. The validation data set is used for model fine-tuning, as well as to assess the adequacy of the model. The test set is used to obtain a final, unbiased estimate of the generalization error of the model. This approach provides a more unbiased estimate of the future performance of the models. However, splitting this small data set containing 355 cases into three parts may result in a loss of information as the number of cases used in building, validating, and testing the models is reduced. Also, the extra sampling process may introduce a new source of variability and decrease the stability of the results (Giudici, 2003). To counter this, however, computer simulation was run for 50 different random generations of the three sets and the paper reports the average, best, and worst error estimates for regression analysis and the ANFIS approach with the standard deviations—all averaged over 50 runs. The most cumulative average error estimates stabilized after about 30 runs.

The subtractive clustering method, a fast one-pass algorithm for estimating the number of clusters and the cluster centers in a set of data, has been used to create the initial neuro-fuzzy inference model. Then the hybrid method was used to optimize the model. With the hybrid method, the forward pass through the data updates the linear parameters using the least squares estimator. In the backward pass, error derivatives are calculated for each node starting from the output end and propagating towards the input end of the network. The final model has two rules. The variables in each rule are connected with the AND logical operator. Each variable has two Gaussian membership functions. Separate computer simulations were run for the system using two rules and three rules. The two-rule simulation generated two membership functions for each variable and the three-rule simulation generated three membership functions for each variable. It appears that the system with two rules and two membership functions for each variable produced the lowest RMSE.

Results and Discussion

To examine the results of the study, three measures are used in comparing the different models. The first measure is the Root Mean Squared Error (RMSE), which is defined as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Actual\_Price_i - Estimated\_Price_i)^2}{N}},$$

where N is the number of test cases.

The second measure is the Maximum Absolute Error (MAE), which is defined as follows:

$$MAE = \max_i |(Actual\_Price_i - Estimated\_Price_i)|.$$

The third measure is the Mean Absolute Percentage Error (MAPE), defined as follows:

$$MAPE = \frac{\sum_{i=1}^{N} \left| \frac{Actual\_Price_i - Estimated\_Price_i}{Actual\_Price_i} \right|}{N} \times 100\%,$$

where N is the number of test cases.

Computer simulation was performed for the three scenarios described above. RMSE, MAE, and MAPE were computed for all three scenarios. As discussed earlier, a common approach is to divide the samples into subsets. The holdout method was used in which the training, validation, and test subsets are independent and the error estimate is pessimistic. The sampling process was repeated 50 times by randomly selecting three different subsets for training, validation, and test. The error estimates over the 50 runs was then averaged to obtain more realistic and reliable error estimates of the models. This is an effective approach to handle the problem of sample size. The approach is commonly used in many published studies and is recommended (Kantardzic, 2003; and Tan, Steinbach, and Kumar,

2006). For each simulation scenario, the cumulative average error estimates were measured over the number of runs and they tend to stabilize after about 30 runs.

Exhibit 8 shows the regression model along with the coefficients and associated statistics for run 39 for the All Variables Scenario. Run 39 has been chosen as it is very representative of the other runs after the errors have stabilized. It is noted that the Presence of Central Air, Construction Type, Wall Type, Basement Type, Garage Type, and Neighborhood variables are recorded on the qualitative (binary, nominal, or categorical) scale. For example, the Wall Type variable has three levels (1, 2, or 3), therefore it is represented by two dummy variables named Wall Type (1) and Wall Type (2). Similarly the Garage Type variable requires three dummy binary variables Garage Type (0), Garage Type (1), and Garage Type (2) in order to represent four different types (categories) of garage. Exhibit 9 shows the regression model (with coefficients and statistics) for the PCA scenario.

It is noted that in Exhibit 8 the coefficients for the variables, Central Air, Second Floor Area, and Upper Floor Area, are negative. The Central Air variable is coded

Exhibit 8 | Regression Equation and Statistics for the All Variables Scenario for Run #39

Parameter	Estimate	Error	*t* Value	Pr > |*t*|
Intercept	37,780.8	12,227.5	3.09	0.0025
Neighborhood 1	239.4	1,635.7	0.15	0.8839
Construction Type 1	12,505.5	6,539.5	−1.91	0.0582
Construction Type 2	11,727.9	7,030.3	−1.67	0.0978
Wall Type 1	1,546.2	2,493.9	0.62	0.5364
Wall Type 2	1,310.1	2,348.4	0.56	0.5779
Year built	53.3	159.3	0.33	0.7383
Square footage in the basement	3.44	4.54	0.76	0.4493
Square footage on the first floor	29.57	8.0	3.69	0.0003
Square footage on the second floor	−22.0	22.6	−0.98	0.3312
Square footage in the upper area	−2.4	7.52	−0.32	0.7503
Number of baths	2,126.3	1,592.8	1.33	0.1844
Presence of central air 0	−1,404.4	1,506.0	−0.93	0.3529
Number of fireplaces	1,188.9	3,724.3	0.32	0.7501
Basement type 0	14,738.1	6,653.9	−2.21	0.0286
Basement type 1	6,408.3	5,134.6	1.25	0.2144
Garage type 0	−3,350.0	4,882.8	−0.69	0.4940
Garage type 1	−2,985.4	6,389.6	−0.47	0.6412
Garage type 2	2,006.7	3,314.9	0.61	0.5461
Garage size (number of cars)	1,452.2	3,239.0	0.45	0.6547

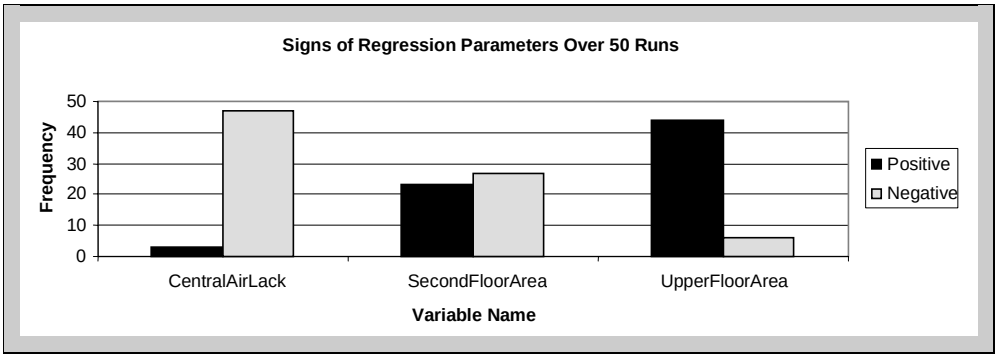
Exhibit 9 | Regression Equation and Statistics for the PCA Scenario for Run #39 that Exhibited the Average Performance on the Test Set—RMSE = $14,186, MAE = $35,241, and MAPE = 16.2%

Parameter	Estimate	Error	*t* Value	Pr > |*t*|
Intercept	45,641.2	6,702.9	6.81	<.0001
PRINCOMP_1	4.277	3.433	1.25	0.215
PRINCOMP_2	7.305	4.178	1.75	0.083
PRINCOMP_3	10.494	5.959	1.76	0.081
PRINCOMP_4	32.824	6.107	5.38	<.0001

as a binary variable that takes the values of 0 or 1 only. The two values indicate lack of central air (0) or presence of central air (1), respectively. SAS treats this variable as a binary variable. In fact, a lack of central air (coded as a 0) decreases the predicted sale price by −$1,404.40 and the presence of central (coded as a 1) increases the predicted sale price by $1,404.40. SAS uses the deviation coding method to treat the binary, nominal, and categorical variables. With deviation coding, the parameters estimate the difference between each level and the average across each level. The parameters for all levels are constrained to sum to zero. This coding is also known as effects coding.

The negative signs for variables Second Floor Area and Upper Floor Area can be explained through a careful examination of the data set. Exhibit 10 shows a histogram of the signs of the three coefficients in the 50 runs. Only 28 and 234 houses (out of 355 houses used in the study) actually have a second floor and an upper area, respectively. That means on the average each of the 50 random training subsets with 142 samples contains about 11 and 94 houses with a second floor and an upper area, respectively. Lack of a second floor or an upper area is denoted

Exhibit 10 | Signs of Selected Regression Coefficients over 50 Runs



by 0s in the data set. It should not be surprising then that the sign for the second floor variable comes out negative 27 times because very few houses in the training subsets actually have a second floor. However, the sign for the upper area variable, which is much better represented in the training subsets, comes out positive 44 times, as shown in Exhibit 10. Similarly, out of 50 runs, the sign for the Presence of Central Air variable is amazingly consistent with the data set and is negative 47 times, which means that a lack of central air in fact decreases the value of the property.

Exhibit 11 reports the summary results for the three scenarios for the test sets as they provide the insight into the true and unbiased future prediction performance of the models. The table shows the averages of RMSE, MAE, and MAPE over the 50 runs. In addition, the table also shows the maximum, minimum, and standard deviation of RMSE, MAE, and MAPE values for all the runs. As can be seen, the results between regression and ANFIS are rather close, although the regression method yielded slightly better results. The difference in RMSE between

Exhibit 11 | Summary Results for the Test Sets for 50 Random Generations

Scenario			RMSE	MAE	MAPE
All Variables	Average	Regression	14,911	44,895	16.8
		ANFIS	15,888	47,246	17.6
	Max.	Regression	17,036	77,784	19.7
		ANFIS	18,606	64,929	20.8
	Min.	Regression	12,975	31,268	14.7
		ANFIS	13,527	34,289	15.1
	Std. Dev.	Regression	968	9,698	1.2
		ANFIS	1,231	7,548	1.3
R^2	Average	Regression	14,267	40,365	16.3
		ANFIS	15,064	43,104	16.7
	Max.	Regression	16,052	51,763	18.9
		ANFIS	15,064	43,104	16.7
	Min.	Regression	12,609	30,218	14.4
		ANFIS	12,735	28,935	14.8
	Std. Dev.	Regression	823	5,028	1.2
		ANFIS	1,120	6,423	1.1
PCA	Average	Regression	14,161	39,911	16.2
		ANFIS	14,432	40,612	16.4
	Max.	Regression	15,814	55,300	19.1
		ANFIS	16,021	52,604	17.9
	Min.	Regression	12,542	29,111	13.6
		ANFIS	12,717	29,602	14.5
	Std. Dev.	Regression	751	5,797	1.2
		ANFIS	766	6,314	0.7

the regression method and ANFIS is 977 for the all variables scenario, 797 for the R^2 variable reduction scenario, and 271 for the PCA scenario. The minimum and maximum RMSE, MAE, and MAPE values also show no clear advantage of one approach to another. It is worth noting that, in general, the difference in average errors improves as the number of variables is reduced, as was expected. For example, the difference in average RMSE between regression and ANFIS goes from 977 to 797 to 271 for the All Variables, R^2, and PCA scenarios, respectively. For MAE, the difference goes from 2,351 to 2,739 to 701. In both RMSE and MAE, the PCA variable reduction method has led to the most dramatic improvement. For MAPE the improvement in difference in average errors is similarly consistent, going from 0.8 for the All Variables scenario to 0.4 for R^2 to 0.2 for PCA. The improvement in the R^2 and PCA scenarios is likely due to the reduced dimensionality in both cases. In the case of R^2, the original 14 input variables have been reduced to 11 and the reduction is much more dramatic in the case of PCA (i.e., from 14 to 4). Variable reduction methods have been shown to be effective and should be considered, especially in cases where the sample size is small. The max and min error values are also very close.

Due to space constraints, only membership functions for two variables are displayed: Wall Type and First Floor Area (see Exhibits 3 and 4 from Run 39). Each of these figures displays the two membership functions corresponding to the associated variable. The linguistic labels Sturdy and Sturdier have been assigned to the Wall Type variable and Small and Large to the First Floor Area variable. As Exhibits 3 and 4 indicate, the membership functions agree with the authors' intuitive knowledge about these two features of a property. The ANFIS has effectively partitioned Wall Type and First Floor Area into two fuzzy sets {Small, Large} for First Floor Area and {Sturdy, Sturdier} for Wall Type. For example, as can be seen from Exhibit 4, the membership of an input value of First Floor Area in the Large fuzzy set increases as the size of the first floor area gets larger. The fuzzy control surface shown in Exhibit 12 shows how more than one input variable may relate to the price. The control surface presents a three-dimensional view of the output (Predicted Sale Price) for two input variables (WallType and FirstFloorArea) while the other 12 input variables are held constant. As can be seen from the figure, the price increases are directly proportional to the first floor area and the sturdiness of the wall type (the higher the value the sturdier the wall). The assignment of the linguistic labels to these variables does not mean that it is always easy to do so. In ANFIS, the coded input values are used in training to create automatic associations between the antecedent coefficients and consequent coefficients. Thus the generated fuzzy rules and membership functions are not always transparent.

In Exhibit 13, the properties are classified into six categories: (1) those with MAPE of less or equal to 5%; (2) those with MAPE between 5% and 10%, (3) those with MAPE between 10% and 15%, (4) those with MAPE between 15% and 20%, (5) those with MAPE between 20% and 25%, and (6) those with MAPE greater than 25%. These ranges are chosen based on the understanding that, say, MAPE $\leq$ 15% may be acceptable and those greater than 15% may be

Exhibit 12 | Control Surface: Price vs. WallType and FirstFloorArea. Other Variables Are Constant. Scenario 1: All Variables, Run 39

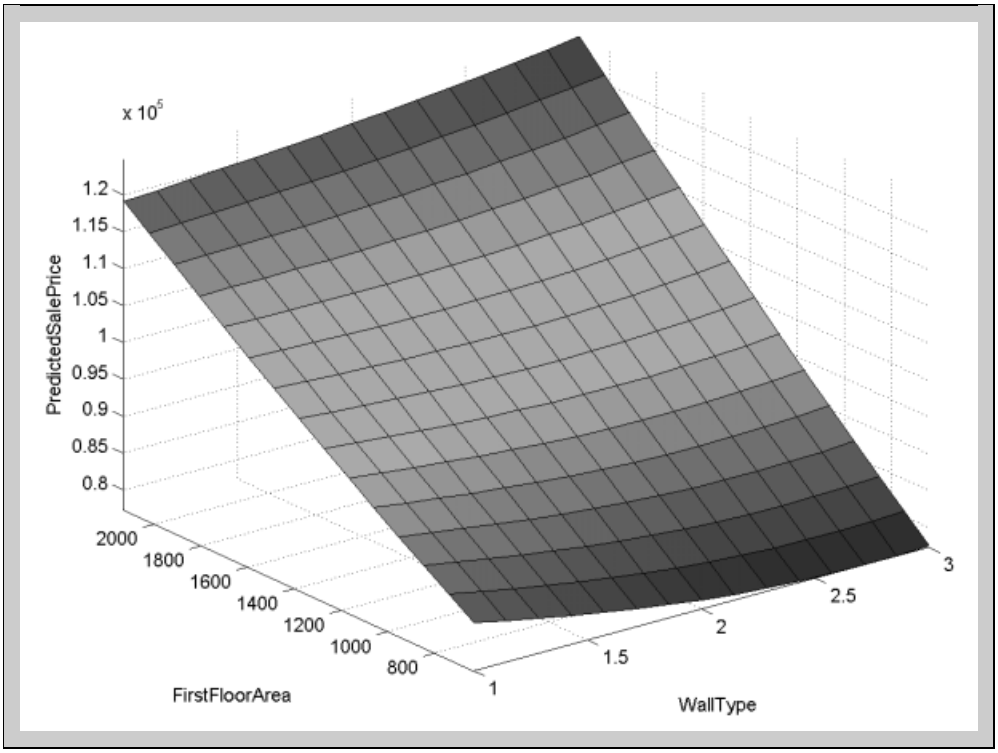


Exhibit 13 | MAPE Percentage Results for the Test Sets Averaged over 50 Random Generations

	Regression Results			ANFIS Results		
Percentage Range	All	R^2	PCA	All	R^2	PCA
≤5%	20.3%	20.2%	20.5%	20.2%	20.3%	19.9%
(5,10%]	18.6%	19.2%	18.1%	18.2%	18.6%	18.3%
(10,15%]	16.5%	17.3%	18.9%	15.9%	16.2%	17.2%
(15,20%]	13.6%	13.5%	14.4%	12.4%	12.6%	13.7%
(20,25%]	9.5%	9.5%	10.1%	9.9%	9.5%	9.9%
>25%	21.5%	20.3%	18.0%	23.4%	22.8%	21.0%

unacceptable. The numbers in the first row (≤5%) show the percentages of test cases that fall into the (≤5%) range. Similarly, the numbers in the sixth row (>25%) show the percentages that fall into the (>25%) range. Out of 106 test cases, regression analysis produces on average 56 cases (52.8%), 57 cases (54.0%), and 58 cases (54.9%) for the All Variables, R^2, and PCA scenarios, respectively, with the cumulative MAPE ≤15%. Similarly, for the three scenarios, the ANFIS system generates 55 cases (51.6%), 55 cases (52.3%), and 56 cases (52.7%), respectively. Again, the results are very close between regression and ANFIS, as observed with RMSE and MAE. It can be noted, however, that the three scenarios of the ANFIS system produce a slightly higher percentage of MAPE (>25%) than the corresponding scenarios in regression analysis (see the last row in Exhibit 13).

Finally, Exhibits 14 and 15 show the antecedent parameters and the consequent parameters for the All Variables scenario and the PCA scenario, respectively. These are the parameters trained or adapted by the ANFIS system. As indicated earlier in the paper, the two-rule simulation yields the best results. Therefore, the total number of fuzzy rules generated is two. Each rule has the following format:

$$\textit{IF } x_1 \textit{ is } A_{1,j} \textit{ AND } x_2 \textit{ is } A_{2,j} \textit{ AND } \cdots \textit{ AND } x_n \textit{ is } A_{n,j}$$
$$\textit{THEN } y = c_0 + c_1x_1 + c_2x_2 + \cdots + c_nx_n,$$

where $A_{i,j}$ is the *jth* linguistic term (such as small, large) of the *i*th input variable x_i and n is the number of inputs (14 for the All Variables scenario and four for the PCA scenario). y is the estimated sale price and c_i are consequent parameters. For simplicity, generic linguistic labels $Small_i$ and $Large_i^2$ are used, as defined by the following membership function:

$$\mu_{A_i}(x) = \exp\left(\frac{-(x - c_i)^2}{2\sigma_i^2}\right).$$

The fuzzy rules are described next, along with the process of fuzzy reasoning, using the PCA scenario. The four input values (principal components) are from Run 39 and are as follows:

PRINCOMP1 = 403.916
PRINCOMP2 = −411.265
PRINCOMP3 = −354.038
PRINCOMP4 = 1,231.694

Exhibit 14 | The ANFIS Rules Consequent Parameters for the All Variables Scenario, Run 39

Variable Name	Rule 1—Output 1 "Predicted Sale Price is Large"	Rule 2—Output 2 "Predicted Sale Price is Small"
Intercept	$1.6295*10^5$	−8,050
Neighborhood	$-3.487*10^4$	$-2.7*10^4$
Construction type	8,690	−7,694
Wall type	9,353	4,505
Year built	−901.2	561.7
Basement area	11.66	−4.934
First floor area	11.02	51.64
Second floor area	24.11	45.84
Upper area	12.5	2.251
Number of baths	−5,456	$-1.05*10^4$
Presence of central air	5,669	−5,503
Number of fireplaces	9,956	399.6
Basement type	2,940	6,280
Garage type	−5,826	$1.9244*10^4$
Garage size (number of cars)	308.3	$-1.4089*10^4$

Exhibit 15 | The ANFIS Rules Consequent Parameters for the PCA Scenario, Run 39

Variable Name	Rule 1—Output 1 "Predicted Sale Price is Large"	Rule 2—Output 2 "Predicted Sale Price is Small"
Intercept	$6.769*10^4$	$4.738*10^4$
PRINCOMP_1	12.25	−11.51
PRINCOMP_2	−9.411	61.17
PRINCOMP_3	−46.38	13.07
PRINCOMP_4	−6.176	37.11

The two fuzzy rules generated are as follows:

Rule 1: If PRINCOMP1 is Large AND PRINCOMP2 is Small AND PRINCOMP3 is Large AND PRINCOMP4 is Large, then Predicted Sale Price is Large.

Rule 2: If PRINCOMP1 is Small AND PRINCOMP2 is Large AND PRINCOMP3 is Small AND PRINCOMP4 is Small, then Predicted Sale Price is Small.

Layer 1 has eight membership functions with their parameters as shown in Exhibit 16. Therefore, given the rules shown, the membership function values are:

$$\mu_{B_{Large}}(403.319) = 0.9823$$

$$\mu_{A_{Small}}(-411.265) = 0.9927$$

$$\mu_{B_{Large}}(-354.038) = 0.6640$$

$$\mu_{B_{Large}}(1{,}231.694) = 0.2883$$

$$\mu_{A_{Small}}(403.916) = 0.7799$$

$$\mu_{B_{Large}}(-411.265) = 0.3863$$

$$\mu_{A_{Small}}(-354.038) = 0.8830$$

$$\mu_{A_{Small}}(1{,}231.694) = 0.1790$$

Layer 2 calculates the weights for the two rules as follows:

$$w_1 = 0.9823*0.9927*0.6640*0.2883 = 0.1867$$

$$w_2 = 0.7799*0.3863*0.8830*0.1790 = 0.0476$$

Exhibit 16 | Membership Functions and Antecedent Parameters

Membership Functions	σ	c
$\mu_{A_{Small}}$(PRINCOMP1)	299.4	192.8
$\mu_{B_{Large}}$(PRINCOMP1)	299.4	347.4
$\mu_{A_{Small}}$(PRINCOMP2)	211.1	−385.7
$\mu_{B_{Large}}$(PRINCOMP2)	211.2	−120
$\mu_{A_{Small}}$(PRINCOMP3)	182.7	−445.2
$\mu_{B_{Large}}$(PRINCOMP3)	182.9	−188.5
$\mu_{A_{Small}}$(PRINCOMP4)	188.6	881.9
$\mu_{B_{Large}}$(PRINCOMP4)	187.8	935.5

Layer 3 normalizes the weights as follows:

$$w_{1\text{-normalized}} = 0.1867/(0.1867 + 0.0476) = 0.7968$$

$$w_{2\text{-normalized}} = 0.0476/(0.1867 + 0.0476) = 0.2032$$

Layer 4 uses the consequent parameters (see Exhibit 15) to compute the sale price for each of the two rules:

Sale Price from the consequent of Rule 1:

$$0.7968*[67{,}690 + 12.25*403.916 + (-9.411)*(-411.265)*(-46.38)*(-354.038) + (-6.176)*1{,}231.694] = \$67{,}984.35$$

Sale Price from the consequent of Rule 2:

$$0.2032*[47{,}380 + (-11.51)*403.916 + (61.17)*(-411.265)*13.07*(-354.038) + 37.11*1{,}231.694] = \$11{,}918.64$$

Finally, Layer 5 computes the final output (Sale Price) as a summation of the results from layer 4:

$$\$67{,}984.35 + \$11{,}918.64 = \$79{,}903$$

Thus, $79,903 represents the final estimated price.

Conclusion

The use of neural networks in property value assessment has attracted considerable interest and has met with varying degrees of success (Do and Grudnitski, 1992; Worzala, Lenk, and Silva, 1995; Guan and Levitan, 1996; McGreal, Adair, McBurney, and Patterson, 1998; Connellan and James, 1998; Bee-Hua, 2000; and Nguyen and Cripps, 2001). More recently fuzzy logic has been proposed as another non-conventional approach to property value assessment (Byme, 1995; and Bagnoli, Smith, and Halbert, 1998). However, one of the main challenges in applying fuzzy logic is the creation of membership functions and fuzzy rules. In simple applications one can build membership functions and fuzzy rules using common sense and/or domain knowledge. In more complex applications where the number of variables is large and the relationships between variables are not

easily discernable, choosing the parameters for a membership function is a trial and error process at best. ANFIS allows creation and refinement of fuzzy rules through neural networks and has received considerable attention in numerous studies in various fields. This paper represents a first attempt to evaluate the feasibility and effectiveness of ANFIS in assessing real estate values.

The results of the study reported in this paper show comparable results to those obtained using a more traditional, multiple regression approach. The results have an average 17.6% error rate (MAPE) when all variables are used. Variable reduction methods R^2 and PCA have been used to help reduce the dimensionality of the data to improve the ANFIS model. They help reduce the error rate to 16.7% in the case of R^2 and 16.4% in the case of PCA (Exhibit 11). The ANFIS model performs slightly worse than the regression model when all the variables are used and the results are very close when variable reduction methods are used.

This study has shown that ANFIS can yield results that are comparable to those obtained using the traditional regression approach. The main contribution of this study is clear demonstration that ANFIS is a viable approach in real estate value assessment and is worthy of further exploration. This study is the first application of the ANFIS to real estate property assessment. The authors hope the results will encourage researchers in this field to further explore the ANFIS approach. The neural networks approach has met with different degrees of success in real estate value assessment. Though there have been some less than desirable results in a few neural networks studies, it has not stopped researchers from continuing to explore that approach. The ANFIS approach holds promise and this paper represents an initial effort and hopefully leads to additional fruitful research in the real estate field.

There are a couple of limitations in this study and these limitations also present further research opportunities. First, although the results of the ANFIS approach are similar to those of the multiple regressions model, it is possible to improve the performance of the ANFIS approach if more cases are available for training. Given the small data set, certain feature domains could be underrepresented because of the small number of training cases and a relatively large number of the input variables. It would be interesting to see if and how a larger data set can improve the ANFIS approach. Another limitation pertains to the types of variables used. In using a file from the county assessor office, the study is limited to the quantifiable variables used in the assessor's multiple regression model. The assessor does not have the ability to deal with more ''fuzzy,'' or nonquantifiable variables. However, the model does have this ability. It could thus use the more emotional, even ''hidden'' input, which directly affects the desirability of a personal residence. Examples include ''the reputation of the neighborhood,'' ''the openness of the interiors,'' etc. The ability to capture and represent these fuzzier features could lead to better assessment of a property's value and fuzzy logic seems to be a natural choice.

Endnotes

[1] We tried different numbers of rules in this study and the system seems to produce the best results when the number of rules is 2.

[2] We have decided to use generic linguistic labels as assigning a semantically meaningful label may not always be easy, especially in the case of the PCA scenario.

References

Bagnoli, C., B. Smith, and C. Halbert. The Theory of Fuzzy Logic and its Application to Real Estate Valuation. *Journal of Real Estate Research*, 1998, 16:2, 169–200.

Bee-Hua, B. Evaluating the Performance of Combining Neural Networks and Genetic Algorithms to Forecast Construction Demand: The Case of the Singapore Residential Sector. *Construction Management and Economics*, 2002, 18:2, 209–18.

Byme, P. Fuzzy Analysis: A Vague Way of Dealing with Uncertainty in Real Estate Analysis. *Journal of Property Valuation & Investment*, 1995, 13:3, 22–41.

Connellan, O. and H. James. Estimated Realization Price by Neural Networks: Forecasting Commercial Property Values. *Journal of Property Valuation & Investment*, 1998, 16:1, 71–86.

Do, Q. and G. Grudnitski. A Neural Network Approach to Residential Property Appraisal. *Real Estate Appraiser*, 1992, 58, 38–45.

Do, Q. and Grudnitski. A Neural Network Analysis of the Effect of Age on Housing Values. *Journal of Real Estate Research*, 1993, 8:2, 253–64.

Giudici, P. *Applied Data Mining: Statistical Methods for Business and Industry*, Chichester, West Sussex, England: John Wiley & Sons, 2003.

Guan, J. and A. Levitan. Artificial Neural Network-Based Assessment of Residential Real Properties: A Case Study. *Accounting Forum*, 1996, 20:3–4, 311–26.

Hasiloglu, A., M. Yilmaz, O. Comakli, and I. Ekmekci. Adaptive Neuro-Fuzzy Modeling of Transient Heat Transfer in Circular Duct Air Flow. *International Journal of Thermal Science*, 2004, 43, 1075–90.

Jang, J.S.R. ANFIS: Adaptive-Network-Based Fuzzy Inference Systems. *IEEE Transactions on Systems, Man, and Cybernetics*, 1993, 23:3, 665–85.

Kantardzic, M. *Data Mining: Concepts, Models, Methods, and Algorithms*. Wiley, 2003.

McGreal, S., A. Adair, D. McBurney, and D. Patterson. Neural Networks: The Prediction of Residential Values. *Journal of Property Valuation & Investment*, 1998, 16:1, 57–70.

Mitchell, T. *Machine Learning*. New York: McGraw-Hill, 1997.

Nguyen, N. and A. Cripps. Predicting Housing Value: A Comparison of Multiple Regression Analysis and Artificial Neural Networks. *Journal of Real Estate Research*, 2001, 22:3, 313–36.

Ponnambalam, K., F. Karray, and S. Mousavi. Optimization Approaches for Reservoir Systems Operation Suing Computational Intelligence Tools. *SAMS*, 2002, 42, 1347–60.

Stepnowski, A., M. Mosynski, and T.V. Dung. Adaptive Neuro-Fuzzy and Fuzzy Decision Tree Classifiers As Applied Seafloor Characterization. *Acoustical Physics*, 2003, 49:2, 189–92.

Sugeno, M. and G.T. Kang. Structure Identification of Fuzzy Model. *Fuzzy Sets and Systems*, 1988, 28: 15–33.

Tan, P-N., M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison-Wesley, 2006.

Worzala, E., M. Lenk, and A. Silva. An Exploration of Neural Networks and its Application to Real Estate Valuation. *Journal of Real Estate Research*, 1995, 10:2, 185–201.

A substantial part of this paper has been revised and resubmitted for review while one of the co-authors was on sabbatical as a Visiting Research Fellow in the School of Computer and Information Science at the Edith Cowan University, Perth, Australia (February–April, 2007). Support from the School is gratefully acknowledged.

Jian Guan, University of Louisville, Louisville, KY 40292 or Jeff.guan@louisville.edu.

Jozef Zurada, University of Louisville, Louisville, KY 40292 or jmzura01@louisville.edu.

Alan S. Levitan, University of Louisville, Louisville, KY 40292 or levitan@louisville.edu.